A NOVEL MODULAR TYPE II RESTRICTION ENDONUCLEASE, CspCI, AND THE USE OF MODULAR ENDONUCLEASES FOR GENERATING ENDONUCLEASES WITH NEW SPECIFICITIES

BACKGROUND OF THE INVENTION

Restriction endonucleases are enzymes that occur naturally in certain unicellular microbes—mainly bacteria and archaea—and that function to protect those organisms from infections by viruses and other parasitic DNA elements. Restriction endonucleases bind to specific sequences of nucleotides ('recognition sequence') in double-stranded DNA molecules (dsDNA) and cleave the DNA, usually within or close to these sequences, disrupting the DNA and triggering its destruction. Restriction endonucleases usually occur with one or more companion enzymes termed modification methyltransferases. Methyltransferases bind to the same sequences in dsDNA as the restriction endonucleases they accompany, but instead of cleaving the DNA, they alter it by the addition of a methyl group to one of the bases within the sequence. This modification ('methylation') prevents the restriction endonuclease from productively recognizing that site thereafter, rendering the site resistant to cleavage. Methyltransferases function as cellular antagonists to the restriction endonucleases they accompany, protecting the cell's own DNA from destruction by its restriction endonucleases. Together, a restriction endonuclease and its companion modification methyltransferase(s) form a

restriction-modification (R-M) system, an enzymatic partnership that accomplishes for microbes what the immune system accomplishes, in some respects, for multicellular organisms.

A large and varied class of restriction endonucleases has been classified as 'Type II' class of restriction endonucleases. These enzymes cleave DNA at defined positions, and when purified can be used to cut DNA molecules into precise fragments for gene cloning and analysis. The biochemical precision of Type II restriction endonucleases far exceeds anything achievable by chemical methods, making these enzymes the reagents *sine qua non* of molecular biology laboratories. In this capacity as molecular tools for gene dissection Type II restriction endonucleases have had a profound impact on the life sciences and medicine in the past 25 years, transforming the academic and commercial arenas, alike. Their utility has spurred a continuous search for new restriction endonucleases, and a large number have been found: today more than 250 Type II endonucleases are known, each possessing different DNA cleavage characteristics (Roberts, R.J. et al., *Nucl. Acids. Res.* 33:D230-D232 (2005)). (Rebase, http://rebase.neb.com/rebase). The production and purification of these enzymes have also been improved by the cloning and overexpression of the genes that encode them, usually in the context of non-native host cells such as *E. coli*.

Since the various restriction enzymes appear to perform similar biological roles, and share the biochemistry of causing dsDNA breaks, it might be thought that they would resemble one

another in amino acid sequence closely. Experience shows this not to be true, however. Surprisingly, far from sharing significant amino acid similarity with one another, most enzymes appear unique, with their amino acid sequences resembling neither other restriction enzymes nor any other known kind of protein. Type II restriction endonucleases seem to have arisen independently of each other during evolution, for the most part, and to have done so hundreds of times, so that today's enzymes represent a heterogeneous collection rather than a discrete family descended from a common ancester. Restriction endonucleases are biochemically diverse in organization and action: some act as homodimers, some as monomers, others as heterodimers. Some bind symmetric sequences, others asymmetric sequences; some bind continuous sequences, others discontinuous sequences; some bind unique sequences, others multiple sequences. Some are accompanied by a single methyltransferase, others by two, and yet others by none at all. When two methyltransferases are present, sometimes they are separate proteins and at other times they are fused. The orders and orientations of restriction and modification genes vary, with all possible organizations occurring. Several kinds of methyltransferases exist, some methylating adenines, others methylating cytosines at the N-4 position, or at the 5 position). Usually there is no way of predicting, *a priori*, which modifications will block a particular restriction endonuclease, which kind(s) of methyltransferases(s) will accompany that restriction endonuclease in any specific instance, nor what their gene orders or orientations will be.

From the point of view of cloning a Type II restriction endonuclease, the great variability that exists among R-M systems means that, for experimental purposes, each is unique. Each enzyme is unique in amino acid sequence and catalytic behavior; each occurs in unique enzymatic association, adapted to unique microbial circumstances; and each presents the experimenter with a unique challenge. Sometimes a restriction endonuclease can be cloned and over-expressed in a straightforward manner but very often it cannot, and what works well for one enzyme may fail altogether for the next. Success with one is no guarantee of success with another.

Novel endonucleases provide opportunities for innovative genetic engineering.

SUMMARY OF THE INVENTION

In an embodiment of the invention, a substantially pure Type IIG restriction endonuclease and an isolated DNA obtainable from *Citrobacter* species 2144 (NEB#1398) (ATCC Patent Accession No. PTA-5846) have been obtained. The recombinant DNA of the enzyme from the *Citrobacter* species and cloned product thereof from *Escherichia coli* NEB#1554 (ATCC Patent Accession No. PTA-5887) is provided.

A further characteristic of the above-described restriction endonuclease is that it recognizes the following base sequence in double-stranded deoxyribonucleic acid molecules:

5'- ↓N_{10}CAANNNNNGTGGN$_{12}$↓ -3' (SEQ ID NO:33)
3'- ↑N_{12}GTTNNNNNCACCN$_{10}$↑ -5' and/or

5'- ↓N_{10}CAANNNNNGTGGN$_{13}$↓ -3' (SEQ ID NO:34)
3'- ↑N_{12}GTTNNNNNCACCN$_{11}$↑ -5' and/or

5'- ↓N_{11}CAANNNNNGTGGN$_{12}$↓ -3' (SEQ ID NO:35)
3'- ↑N_{13}GTTNNNNNCACCN$_{10}$↑ -5' and/or

5'- ↓N_{11}CAANNNNNGTGGN$_{13}$↓ -3' (SEQ ID NO:32)
3'- ↑N_{13}GTTNNNNNCACCN$_{11}$↑ -5'

and cleaves the DNA on both sides of the recognition sequence at the alternative positions shown by the arrows.

The DNA encoding the restriction endonuclease described above may include a first DNA segment expressing endonuclease and methyltransferase catalytic functions and a second DNA segment encoding a sequence specificity function of the restriction endonuclease wherein the first and second DNA segments are contained in one or more DNA molecules.

The above-described DNA may be inserted into a vector. The vector may include at least one of a first DNA segment coding for the restriction and modification domains of CspCI

restriction endonuclease and a second segment coding for the specificity domain of the restriction endonuclease.

In an embodiment of the invention, a host cell is provided which is transformed by a first DNA segment coding for the restriction and modification domains of CspCI restriction endonuclease and a second segment coding for the specificity domain of the restriction endonuclease. The first DNA segment and the second DNA segment may be contained within one or more DNA vectors.

In an embodiment of the invention, a method is provided for obtaining the restriction endonuclease which includes the steps of cultivating a sample of *Citrobacter* species 2144 (NEB#1398) or a host cell as described above under conditions favoring the production of the endonuclease; and purifying the endonuclease therefrom.

In an embodiment of the invention, a method of making a Type II restriction endonuclease having an altered specificity includes: (a) selecting a restriction endonuclease from a set of enzymes wherein each enzyme in the set is characterized by a modular structure having a specificity subunit and a catalytic subunit. The specificity subunit further includes an N-terminal domain for binding one half site of a bipartite recognition sequence and a C-terminal domain for binding a remaining half site of the bipartite recognition sequence; (b) modifying the

specificity subunit; and (c) obtaining the restriction endonuclease with altered specificity.

Where the restriction endonuclease is CspCI, one half site is CAA and the other half site is GTGG.

In this method, the step of modifying the specificity subunit may further include (a) substituting the N-terminal domain with a second copy of the C-terminal domain or substituting the C-terminal domain with a second copy of the N-terminal domain (b) substituting the N-terminal domain or the C-terminal domain or both N-terminal and C-terminal domain with a DNA-binding domain from a second restriction endonuclease or methylase, or (c) mutating the N-terminal domain, the C-terminal domain or both domains to alter the binding specificity. In any of these modifications or without these modifications, an additional modification can be added, namely changing the length of the spacer amino acid sequence between the N-terminal and C-terminal domains of the specificity subunit. In any of the above, the specificity subunit and the catalytic subunit may be encoded by separate and distinct genes.

In an embodiment of the invention, DNA-binding domain from the second restriction endonuclease or methylate may derive from a Type I restriction endonuclease, another Type IIG restriction endonuclease, or from a γ-type m^6A methyltransferase. Additionally, it is envisioned that the N-

terminal cleavage domains can be grafted onto other DNA-binding proteins.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 is an agarose gel showing CspCI-cleavage of phage lambda, T7, PhiX174, pBR322 and pUC19 DNAs. Lanes are as follows:

lanes 1, 10, 15: lambda-HindIII, PhiX174-HaeIII size standards;

lane 2: lambda DNA + CspCI;

lane 3: T7 DNA + CspCI;

lane 4: PhiX174 DNA;

lane 5: PhiX174 DNA + CspCI;

lane 6: PhiX174 DNA + CspCI + PstI;

lane 7: PhiX174 DNA + CspCI + SspI;

lane 8: PhiX174 DNA + CspCI + NciI;

lane 9: PhiX174 DNA + CspCI + StuI;

lane 11: pBR322 DNA;

lane 12: pBR322 DNA + CspCI;

lane 13: pUC19 DNA;

lane 14: pUC19 DNA + CspCI.

Figure 2 is a high-concentration agarose gel of CspCI-cleaved pUC2CspC DNA showing 35±1 bp internal 'mini-fragment' (arrows).

Figure 3 is a high-resolution agarose gel showing partial-digestion doublet fragments. DNA: BglI-cleaved pUC2CspC re-digested with increasing amounts of CspCI. Transient CspCI-BglI fragment doublets are show by the arrows.

Figures 4a and 4b show a determination of the CspCI cleavage sites by primed synthesis. Two experiments were performed using the same M13mp18 template and primer combination. (-) is CspCI-cleaved DNA only; (+) is Klenow-treatment of the CspCI-cleaved DNA.

Figure 5 shows a determination of the CspCI cleavage sites by run-off automated sequencing.

Figure 5a: pUC1CspC-4 template; forward primer (SEQ ID NO:1)

Figure 5b: pUC1CspC-4 template; reverse primer (SEQ ID NO:2)

Figure 5c: pUC1CspC-1 template; forward primer (SEQ ID NO:3)

Figure 5d: pUC1CspC-1 template; reverse primer (SEQ ID NO:4

A-anomalies, signifying template cleavage, are shown as triangles (Δ) below the tracings.

Figure 6 shows the complete nucleotide sequence of the DNA cloned from *Citrobacter* species 2144 (NEB#1398, New England Biolabs, Inc., Beverly, MA) (SEQ ID NO:5).

Figure 7a shows the nucleotide sequence of the CspCI-R-M gene (SEQ ID NO:6).

Figure 7b shows the nucleotide sequence of the CspCI-S gene (SEQ ID NO:7).

Figure 8a shows the gene organization of the CspCI restriction-modification system.

Figure 8b shows the gene organization of the plasmid clone pUC19-CspCI-R-M-S ApoI #3 carrying the CspCI genes inserted into the EcoRI site of pUC19

Figure 9a shows the predicted amino acid sequences of the R-M-CspCI endonuclease-methyltransferase subunit (SEQ ID NO:8).

Figure 9b shows the predicted amino acid sequences of the CspCI specificity subunit (SEQ ID NO:9).

DETAILED DESCRIPTION OF THE INVENTION

In most restriction enzymes, the parts of the protein responsible for binding to the recognition sequence ('specificity':S) and for cleaving it ('catalysis') are interlinked. Experience has taught that altering either of these functions frequently impairs the other, and renders the enzyme inactive. A new class of enzymes has been identified in which the functions of specificity and catalysis are largely separated. These members of the Type IIG class of restriction endonucleases are large enzymes in which the twin activities of restriction and modification are combined in a single polypeptide chain while specificity resides with a different polypeptide chain. Examples of restriction endonucleases in this class are CspCI, BcgI and BaeI. While not wishing to be limited by theory, CspCI is believed to act as a dimer of one R-M-subunit and one S-subunit, while BcgI acts as a trimer of two R-M subunits and one S-subunit.

The separated functional organization of this class of enzymes provides unusual opportunities for protein engineering because the functional modules can be independently manipulated to generate novel specificities of choice as described in more detail in Example V.

This new class of endonucleases is characterized by a DNA encoding the specificity subunit that is distinct from the R-M

genes. The genes for these occur side by side, naturally, and are expressed in *cis*. These genes can also be separated into different replicons, and expressed in *trans*, without loss of activity. The separate location of these genes in different amplicons permits the S and the R-M genes to be altered individually, and allows the endonuclease, or variants of it, to be reconstituted easily *in vivo*, simply by introducing the two replicons into the same cell, rather than rejoining the genes into the same DNA molecule. Reconstitution can be performed individually, or in bulk by transforming libraries of one altered gene into cells harboring the other. Both genes may alternatively be co-transformed, together in a mixture.

Alternatively, the R-M and S genes can be separated to allow them to be expressed individually in different host cells. It will be appreciated that since neither protein alone exhibits toxic activity, the cells producing either subunit will be viable. Expressing the subunits separately allows them to be purified individually, and enables the enzyme, or variants of it, to be reconstituted easily *in vitro*, simply by mixing together preparations of the two subunits. High-throughput screening, and/or multiplexing can be achieved using extracts of cells instead of purified proteins.

The presence of DNA-methyltransferase motifs within this class of endonuclease suggests that the endonucleases have intrinsic methylation activity, in addition to endonuclease

activity. For example, CspCI is dependent on S-adenosyl-L-methionine (AdoMet). By mutating the catalytic sites for these activities, variants of these endonucleases can be isolated. DNA-cleavage activity, DNA-methylation activity, or both, may be abolished in these mutants.

Typically, the specificity subunit of endonucleases in the Type IIG class determines which target sequence in a DNA molecule will undergo cleavage by means of the R-M subunit. The R-M subunit has a distinct N-terminal domain for DNA-cleavage, and a distinct C-terminal domain for DNA-methylation. The S subunit has a distinct N-terminal domain for binding one-half of the bipartite recognition sequence, and a distinct C-terminal domain, for binding the other half.

Other modular enzymes exist which characteristically cleave DNA at a sequence that is distant to the recognition site. However, these enzymes are monomers (CjeI and AloI) or homodimers (HaeIV) both types being single proteins with a composition of R-M-S.

For any unknown restriction endonuclease that is observed to have a modular structure, the recognition sequence of the endonuclease of the class may be determined by mapping the locations of the cleavage sites in a target DNA of known sequence. The DNA sequences of these regions are compared for similarity and common features. Candidate recognition sequences are compared with the observed restriction fragments

produced by endonuclease-cleavage of a variety of DNAs. The approximate size of DNA fragments produced by endonuclease digestion can be entered into the program REBPredictor, which can be accessed at http://tag.neb.com/~vincze/REBpredictor/index.php. Example III describes how REBPredictor was used to predict potential recognition sites for CspCI.

A modular endonuclease of the type described above can be obtained as a product of recombination in a host cell or by culturing the native strain. Host cells are grown in suitable media supplemented with 100mg/ml ampicillin and incubated aerobically at 37°C. Cells in the late logarithmic stage of growth are collected by centrifugation and either disrupted immediately or stored frozen at -70°C.

Conventional protein purification techniques can be used to isolate the endonuclease from lysed cells. Cell paste is suspended in a buffer solution and ruptured by sonication, high-pressure dispersion or enzymatic digestion to allow extraction of the endonuclease by the buffer solution. Intact cells and cellular debris are then removed by centrifugation to produce a cell-free extract containing the endonuclease. The endonuclease is then purified from the cell-free extract by ion-exchange chromatography, affinity chromatography, molecular sieve chromatography, or a combination of these methods.

Alteration of the specificity domains in Type I restriction enzymes has been achieved to generate novel enzymes that recognize symmetric DNA sequences, and hybrid DNA sequences (Bickle et al. *Journal of Cell Biochemistry* 18c136 (1994); Bickle et al. *EMBO Journal* 15: 4775-4783 (1996)). Example VI describes how the specificity domain in a modular Type II restriction enzyme can be manipulated to alter the specificity of the enzyme.

Present embodiments of the invention are further illustrated by the following Examples. These Examples are provided to aid in the understanding of embodiments of the invention and are not construed as a limitation thereof.

The references cited above and below as well as provisional application number 60/555,795 are herein incorporated by reference.

EXAMPLES

Example I: Isolation of CspCI

CspCI was obtained by culturing either (i) *Citrobacter* species 2144 (NEB#1398) or (ii) the transformed host, *E. coli* NEB#1554, and recovering the endonuclease from the cells. A sample of *Citrobacter* species 2144 (NEB#1398) has been deposited under the terms and conditions of the Budapest Treaty with the American Type Culture Collection (ATCC) on March 4,

2004 and bears the Patent Accession No. PTA-5846. A sample of a recombinant strain expressing CspCI, *E. coli* (NEB#1554), has also been deposited under the terms and conditions of the Budapest Treaty with the American Type Culture Collection (ATCC) on March 24, 2004 and bears the Patent Accession No. PTA-5887.

Citrobacter species 2144 (NEB#1398) or *E. coli* (NEB#1554) were incubated aerobically at 37°C. Cells in the late logarithmic stage of growth are collected by centrifugation and either disrupted immediately or stored frozen at -70°C.

The CspCI endonuclease was isolated from *Citrobacter* species 2144 (NEB#1398) or *Escherichia coli (*NEB#1554) by conventional protein purification techniques. The cell paste was suspended in a buffer solution and ruptured by sonication, high-pressure dispersion or enzymatic digestion to allow extraction of the endonuclease by the buffer solution. Intact cells and cellular debris are then removed by centrifugation to produce a cell-free extract containing CspCI. The CspCI endonuclease was then purified from the cell-free extract by ion-exchange chromatography, affinity chromatography, molecular sieve chromatography, or a combination of these methods to produce the endonuclease.

Example II: Production of Native or Recombinant CspCI Endonuclease

277 grams of *E. coli* NEB#1554 CspCI cell pellet or *Citrobacter* species 2144 (NEB#1398) (New England Biolabs, Inc., Beverly, MA) were suspended in 1 liter of Buffer A (20mM Tris-HCl (pH 7.4), 1.0mM DTT, 0.1mM EDTA, 5% Gycerol) containing 300mM NaCl, and passed through a Gaulin homogenizer at ~12,000 psig. The lysate was centrifuged at ~13,000 x G for 40 minutes and the supernatant collected.

The supernatant solution was applied to a 400 ml DEAE Fast-Flow column (GE Healthcare, formerly Amersham Biosciences, Piscataway NJ) column equilibrated in buffer A plus 300mM NaCl, and the flow-through, containing the CspCI endonuclease activity, was diluted 1:1 with buffer A.

The diluted enzyme was applied to a 375 ml Heparin Hyper-D column (Biosepra, Marlborough MA), which had been equilibrated in buffer B. (20mM Tris-HCl (pH 7.4), 150mM NaCl, 1.0mM DTT, 0.1mM EDTA, 5% Gycerol). A 2.5 L wash of buffer B was applied, then a 2 L gradient of NaCl from 0.15M to 1M in buffer B was applied and fractions were collected. Fractions were assayed for CspCI endonuclease activity by incubating with 1 microgram of phage lambda DNA (NEB) in 50 microliter NEBuffer 2, supplemented with 20 microMolar (AdoMet) for 15 minutes at 37° C. CspCI activity eluted at 0.3M to 0.35M NaCl.

The Heparin Hyper-D column fractions containing the CspCI activity were pooled and load directly onto a 200 ml Ceramic htp column (Biosepra, Marlborough MA) equilibrated in

Buffer B. A 1 L wash of buffer B was applied, then a 1 L gradient of $KHPO_4$ (pH 7.5) from 0M to 0.6M in buffer B was applied and fractions were collected. Fractions were assayed for CspCI endonuclease activity by incubating with 1 microgram of phage lambda DNA in 50 microliter NEBuffer 2, supplemented with 20 microMolar AdoMet for 15 minutes at 37° C. CspCI activity eluted at 0.4M to 0.5M KHPO4.

The Ceramic HTP column fractions containing the CspCI activity were pooled and dialyzed into Buffer C (20mM Tris-HCl (pH 7.4), 100mM NaCl, 1.0mM DTT, 0.1mM EDTA, 5% Gycerol).

This pool was flowed through a 50 ml Source Q column (GE Healthcare, formerly Amersham Biosciences, Piscataway NJ.) equilibrated in buffer C and directly onto a Heparin TSK equilibrated in buffer C. A 250 ml wash of buffer C was applied, then a 400 ml gradient of NaCl from 0.1M to 0.8 M in buffer C was applied and fractions were collected. Fractions were assayed for CspCI endonuclease activity by incubating with 1 microgram of phage lambda DNA (New England Biolabs, Inc., Beverly, MA) in 50 microliter NEBuffer 2, supplemented with 20 microMolar AdoMet for 15 minutes at 37° C. CspCI activity eluted at 0.3M to 0.35M NaCl.

The pool was dialyzed into Storage Buffer (10mM Tris-HCl (pH 7.4), 100mM NaCl, 1.0mM DTT, 0.1mM EDTA, 50% Gycerol). One million units of CspCI were obtained from this procedure. The CspCI endonuclease thus produced was

substantially pure and free of contaminating nucleases. SDS polyacrylamide gel electrophoresis of a sample of this preparation showed it comprised two principal proteins of approximately 70 kDa and 35 kDa in the approximate ratio by mass of 2:1.

Activity determination

CspCI activity: Samples of from 1 to 10 microliter were added to 50 microliter of substrate solution consisting of 1X NEBuffer 2 (New England Biolabs, Inc., Beverly, MA) containing 1 microgram of phage lambda phage DNA supplemented with 20 microMolar AdoMet. The reaction was incubated at 37°C for 60 minutes. The reaction was terminated by adding 20 microliter of stop solution (50% glycerol, 50 mM EDTA pH 8.0, and 0.02% Bromophenol Blue.) The reaction mixture was applied to a 1% agarose gel and electrophoresed. The bands obtained were identified by comparison with DNA size standards.

Unit Definition: One unit of CspCI is defined as the amount of CspCI required to completely cleave one microgram of phage lambda DNA in a reaction volume of 50 microliter of 1X NEBuffer 2 (New England Biolabs, Inc., Beverly, MA) supplemented with 20 microMolar AdoMet, within one hour at 37°C.

Properties of CspCI:

AdoMet: Supplementing the CspCI reaction with 20 mM AdoMet greatly enhanced the activity of the enzyme. In reactions where AdoMet was omitted, the enzyme exhibited less than 5% of the cutting activity it exhibited in the AdoMet-supplemented reactions, indicating that AdoMet is a necessary cofactor for this enzyme.

Activity in various reaction buffers: CspCI was found to be most active in NEBuffer 2 + AdoMet, relative to other standard NEBuffers (New England Biolabs, Inc, Beverly, MA).

Digestion at 37°C for one hour in the following NEBuffers yielded the following approximate percentage cleavage activities relative to NEBuffer 2 (New England Biolabs, Inc, Beverly, MA)+ 20mM AdoMet:

NEBuffer 1 + 20mM AdoMet: 10%
NEBuffer 2 + 20mM AdoMet: 100%
NEBuffer 3 + 20mM AdoMet: 10%
NEBuffer 4 + 20mM AdoMet: 75%
NEBuffer 2 – (No AdoMet): < 5%

Activity in a 16-hour reaction: O.5 units of CspCI are required to cut one microgram of phage lambda DNA in a 16-hour digest, compared to one unit that is required to cut one microgram of phage lambda DNA in a one-hour digest.

Temperature: The CspCI unit titer was determined at 37°C by a one-hour incubation in 1X NEBuffer 2 plus 20 microMolar AdoMet. Incubation of CspCI at 70°C for 20 minutes prior to performing a reaction at 37°C does not inactivate the enzyme. After heat treatment at 70°C for 20 minutes, CspCI retains nearly full activity.

Bilateral cleavage: CspCI cleaves DNA on both sides of its recognition sequence. As a result, in addition to producing regular restriction fragments, CspCI cleavage generates small, internal, 'mini-fragments' of 35±1 bp, one from each recognition site. These mini-fragments, which can be visualized by gel electrophoresis (Figure 2), comprise the recognition sequence and the flanking DNA on each side up to the cut sites. The two cleavage events that produce the mini-fragments appear to proceed separately: cleavage occurs first on one side of the recognition sequence and then later on the other side, rather than on both sides simultaneously. As a result, when partially digested samples of DNA are examined by gel electrophoresis, the DNA fragments appear as doublets or triplets depending on whether the min-fragments have been trimmed yet from their termini (Figure 3).

Example III: Determination of the CspCI Cleavage Site

The location of CspCI-induced cleavage relative to the recognition sequence was determined by two methods, primed synthesis and run-off automated sequencing.

A: Primed synthesis method

The locations of CspCI cleavages relative to the recognition sequence was determined by cleavage of a primer extension product, which was then electrophoresed alongside a set of standard dideoxy sequencing reactions produced from the same primer and template. M13mp18 DNA was employed as the template with a primer near the recognition sequence position at 3009. Readable sequence for this primer template combination begins at position 3069 and continues through the CspCI site.

Sequencing Reactions

The sequencing reactions were performed using the Sequenase version 2.0 DNA sequencing kit (GE Healthcare, formerly Amersham Life Science) with modifications for the cleavage site determination. The template and primer were assembled in a 0.5 ml Eppendorf tube by combining 2.5

microliter dH2O, 3 microliter 5X sequencing buffer (200 mM Tris pH 7.5, 250 mM NaCl, 100 mM $MgCl_2$), 8 microliter M13mp18 single-stranded DNA (1.6 microgram) and 1.5 microliter of primer at 3.2 mM concentration. The primer-template solutions were incubated at 65°C for 2 minutes, then cooled to 37°C over 20 minutes in a beaker of 65°C water on the bench top to anneal the primer. The labeling mix (diluted 1:20) and T7 Sequenase polymerase were diluted according to manufacturer's instructions. The annealed primer and template tube was placed on ice. To this tube were added 1.5 microliter 100mM DTT, 3 microliter diluted dGTP labeling mix, 1 microliter [α-^{33}P] dATP (2000Ci/mM, 10mCi/ml) and 3 microliter diluted T7 Sequenase polymerase (GE Healthcare, formerly Amersham, Piscataway, NJ). The reaction was mixed and incubated at room temperature for 4 minutes.

3.5 microliter of this reaction was then transferred into each of four tubes containing 2.5 microliter termination mix for the A, C, G and T sequencing termination reactions. To the remaining reaction was added to 10 microliter of Sequence Extending Mix (GE Healthcare, formerly Amersham Biosciences, Piscataway, NJ), which is a mixture of dNTPs (no ddNTPs) to allow extension of the primer through and well beyond the CspCI site with no terminations to create a labeled strand of DNA extending through the CspCI recognition site for subsequent cleavage. The reactions were incubated 5 minutes at 37°C. To the A, C, G and T reactions were added 4 microliter of stop solution and the samples were stored on ice. The extension

reaction was then incubated at 70°C for 20 minutes to inactivate the DNA polymerase (Sequenase) (GE Healthcare, formerly Amersham, Piscataway, NJ), then cooled on ice.

10 microliter of the extension reaction was then placed in zone 0.5 ml Eppendorf tube and 7 microliter was placed in a second tube. To the first tube was added 1 microliter (approximately 0.5 unit) of CspCI endonuclease, The reaction was mixed, and then 2 microliter was transferred to the second tube. These enzyme digest reactions were mixed and then incubated at 37°C for 1 hour, following which the reactions were divided in half. To one half, 4 microliter of stop solution was added and mixed (the 'minus' polymerase reaction). To the second half, 0.4 microliter Klenow DNA polymerase (NEB#210) (New England Biolabs, Inc., Beverly, MA) containing 80 mM dNTPs was added (the 'plus' reaction), and the reaction was incubated at room temperature for 15 minutes, following which 4 microliter of stop solution was added.

The sequencing reaction products were electrophoresed on an 6% Bis-Acrylamide sequencing gel (Stratagene Corporation, La Jolla, CA), with the CspCI digestions of the extension reaction next to the set of sequencing reactions produced from the same primer and template combination.

Results

Digestion of the extension reaction product (the 'minus' reaction) produced a band which co-migrated with the C residue 12 bases 5' to the CspCI recognition sequence, 5'-CAGAGAGATAACCCACAAGAATTG-3', (SEQ ID NO:10) indicating cleavage between the 12^{th} and 11^{th} bases 5' of the recognition sequence on this strand. A second band was produced which co-migrated with the A residue 12 bases 3' to the CspCI recognition site on this strand, CCACAAGAATTGAGTTAAGCCCAA (SEQ ID NO:11), indicating cleavage between the 12^{th} and 13^{th} bases 3' to the recognition site. There was also a faint band one base farther from the recognition site, indicating that a small portion of the molecules were cut between the 13^{th} and 14^{th} bases 3' to the recognition sequence. Treatment of the cleaved extension reaction product with Klenow DNA polymerase (the 'plus' reaction) produced a band two bases shorter than the first band described above, which co-migrated with the A residue 14 bases 5' to the recognition sequence; 5'-ATCGAGAGATAACCCACAAGAATTG-3' (SEQ ID NO:12), indicating cleavage between the 13^{th} and 14^{th} bases 3' to the recognition sequence on the opposite strand of the DNA (5'-CAANNNNNGTGG(N_{13}) (SEQ ID NO:13). Several additional bands were observed in the 'plus' lane as well, corresponding to the original band, 12 bases 3' to the site, and bands one and two bases shorter, produced from cuts on the opposite strand of DNA closer to the recognition sequence (Figure 4).

These results, when combined with those obtained by the second method described below, indicate that CspCI cleaves DNA on both sides of its recognition sequence, and can do so at either N11/N13 or N10/N12 5′ to the sequence 5′-CAANNNNNGTGG-3′ (SEQ ID NO:14) and at N13/N11 or N12/N10 3′ to the sequence, to produce DNA fragments with 2-base 3′-extensions, and an excised fragment of 34, 35 or 36 bases that contains the recognition site.

B: Run-off sequencing method

The second approach employed automated sequencing of CspCI-partially cleaved template DNA with forward and reverse primers to produce sequencing traces that extended through the sites of cleavage. Two plasmids served as templates, pUC1CspC-1 and pUC1CspC-4, constructed by inserting an oligonucleotide containing the CspCI recognition sequence into the AatII site at nt 2617 of pUC19 in both orientations (described in Example III, section 2, below).

<u>CspCI-cleavage of pUC1CspC-1 and pUC1CspC-4</u>

Sequencing reactions were carried out on partial digests of pUC1CspC-1 and pUC1CspC-4, in order to determine the sites of cleavage on both sides of the recognition site.

The digests were performed as follows:

a. Combine:

25 microgram pUC1CspC-1 or pUC1CspC-4
100 microliter NEBuffer2
1 microliter 32 mM AdoMet
dH2O to 1000 microliter

b. Distribute the mixture: 200 microliter in one reaction tube, 100 microliter in 8 subsequent tubes.

c. Add 160 units CspCI endonuclease to the first tube, mix, remove 100 microliter and add it to the second tube, mix, remove 100 microliter and add it to the third tube, etc. until the 9th tube is reached.

d. Incubate all 9 reactions at 37°C for 60 minutes, then place on ice.

e. Analyze a sample of each reaction on agarose gel; select completely cleaved and partially cleaved plasmids.

f. Purify the cleaved plasmids for sequencing using Zymo DNA Clean and Concentrator-5 spin-columns according to the manufacturer's recommendations (Zymo Research, Orange, CA).

Sequencing Reactions

The reactions were performed with an ABI377 DNA sequencer using CspCI-cleaved pUC1CspC-1 and -4 plasmid

templates, and a pair of primers that initiate synthesis approximately 250 nt away from the CspCI site on one side, (forward-primer), and 160 nt away from the CspCI site on the other side (reverse primer). The sequences of these two primers are:

5'- CAGTTCGATGTAACCCACTCG -3' (SEQ ID NO:15) forward primer; corresponds to pUC19 nt 2346-2366; interrogates the minus-strand of the vector.

5'- CCCGCTGACGCGCCCTGACGGGC -3' (SEQ ID NO:16) reverse primer; corresponds to pUC19 nt 96-118 complement; interrogates the plus-strand of the vector.

When sequencing reactions encounter the 5′ end of a template strand, they frequently add a final, non-templated A to the synthesized strand. If the template DNA comprises a mixture of intact and truncated strands, such as occurs in incompletely cleaved DNA samples, the position of cleavage reveals itself in the sequencing trace by an anomalous A peak superimposed on the normal peak, and by an overall reduction in the heights of the following peaks. If the base normally present at the position of the anomaly is something other than A — G, for example — then a mixed signal is seen, in this example G plus A. However, if the base normally present at this position is also A, then a single A peak is seen, perhaps higher than normal, and this confounds unambiguous identification.

Results

Unambiguous results were obtained for the positions of cleavage on the 5′ sides of the recognition sequence, but the data was poorer regarding cleavage on the 3′ sides. As a whole, however, they were consistent with the endonuclease cleaving to produce fragments with 2-base 3′-overhangs at. Sequence traces from representative reactions are shown in Figure 5.

The reaction of partially cleaved pUC1CspC-4 with the forward primer displayed a strong anomalous A superimposed on the G 13 nt before the recognition sequence, and a stronger-than-expected A peak 11 nt after it:

5′...AAGTGccacctgacgtgcaacctaggtggcacgtctaagaaac... (SEQ ID NO:17)

(Notation. Underlined: CspCI recognition site; bold: normal base over which anomalous A superimposed; UPPER CASE: peaks of normal height; lower case: peaks of reduced height)

These results suggest that cleavage of the complementary strand (indicated |) occurs:

5′...GTTT|CTTAGACGTGCCACCTAGGTTGCACGTCAGGTGGC |ACTT... (SEQ ID NO:18)

The reaction of partially cleaved pUC1CspC-4 with the reverse primer displayed a strong A-anomaly on the T 12 nt

before the recognition sequence, and a suggestion of two anomalous A's under the two G's 11 and 12 nt after the sequence:

5'...TGGTTtcttagacgtgccacctaggttgcacgtcaggtggcact...

(SEQ ID NO:19)

Ignoring momentarily the G-11 anomaly, these results suggests that cleavage of the complementary strand occurs:

5'...TGC|CACCTGACGTGCAACCTAGGTGGCACGTCTAAGAA|ACCA...

(SEQ ID NO:20)

Combining these results, CspCI-cleavage at the site in pUC1CspC-4 appears to be:

5'...AGTGC|CACCTGACGTGCAACCTAGGTGGCACGTCTAAGAA|ACC...

(SEQ ID NO:21)

3'...TCA|CGGTGGACTGCACGTTGGATCCACCGTGCAGATTC|TTTGG...

(SEQ ID NO:22)

That is to say: 11/13 CAA N_5 GTGG 12/10 (SEQ ID NO:14)

The same G-13 and A-11 A-anomalies were seen when partially-cleaved pUC1CspC-1 was interrogated the forward primer, and the same T-12 A-anomaly was seen when it was interrogated with the reverse primer. Consequently, cleavage at the site in pUC1CspC-1 appears to be:

5'...AGTGC|CACCTGACGTGCCACCCGGGTTGCACGTCTAAGA A|ACC...

(SEQ ID NO:23)

3'...TCA|CGGTGGACTGCACGGTGGGCCCAACGTGCAGATTC| TTTGG...

(SEQ ID NO:24)

That is to say: 10/12 CAA N_5 GTGG 13/11

(SEQ ID NO:14)

This numerical reversal in cleavage distances indicates that the positions of DNA cleavage are independent of recognition-sequence orientation, and dependent on nature of flanking sequence. The sequence to the left (counter-clockwise) of the recognition site is the same in both plasmids, as also is the sequence to the right (clockwise). The latter, which is somewhat A:T-rich, would seem to be more extended, physically, than the G:C-rich DNA to the left, such that the endonuclease, as it 'measures' out from its binding site, cleaves 12/10 on either side if the DNA is extended, and 13/11 on either side if the DNA is compact.

Returning to the G-11 anomaly momentarily ignored, above, its presence in the pUC1CspC-4/reverse primer reaction suggests that the otherwise compact leftward DNA can become more extended, perhaps due to torsional relaxation that accompanies supercoil-release during digestion, leading to

10/12 cleavage at this location, also. This confirms to a degree that CspCI can also cleave:

10/12 CAA N_5 GTGG 12/10 (SEQ ID NO:14), and by extension,

11/13 CAA N_5 GTGG 13/11 (SEQ ID NO:14).

Example IV: Cloning of the CspCI Restriction-Modification Genes

1. Preparation of genomic DNA

Genomic DNA was prepared from 2.5g of *Citrobacter* species 2144, by the following steps:

a. Cell wall digestion by addition of lysozyme (2 mg/ml final), sucrose (1% final), and 50 mM Tris-HCl, pH 8.0.

b. Cell lysis by addition of 24 ml of Lysis mixture: (50mM Tris-HCl pH 8.0, 62.5mM EDTA, 1% Triton.

c. Removal of proteins by phenol-$CHCl_3$ extraction of DNA 2 times (equal volume).

d. Dialysis in 4 liters of TE buffer, buffer change four times.

e. RNase A treatment to remove RNA.

f. Genomic DNA precipitation in 0.4M NaCl and 0.55 volume of 100% isopropanol, spooled, dried and resuspended in TE buffer.

2. Preparation of plasmid vector pUC2CspC

Plasmid cloning vector pUC2CspC was constructed from E.coli cloning vector pUC19 by inserting two CspCI recognition sites, one at the unique AatII site at nt 2617, and another at the DraI site at nt 1563.

a. Two pairs of complementary oligonucleotides were synthesized. Annealing of each pair produces a CspCI recognition site, and double-stranded ends that can be ligated to either AatII or DraI DNA fragments such that the ligation product no longer contains the AatII or DraI site.

The oligonucleotide sequences, shown below in annealed double-strand format, were:

AatII-site linker:

```
          5'-GCAACCNGGGTGGCACGT-3'
               ||||||||||||||||
3'-TGCACGTTGGNCCCACCG-5'
```

(SEQ ID NO:25)

DraI-site linker:

5'-CAANNNNNGTGG-3'
||||||||||||
3'-GTTNNNNNCACC-5'

(SEQ ID NO:14)

b. For the AatII site linker, 1 microgram pUC19 was digested in a small volume with AatII.

c. Annealed oligonucleotide linker was added to the reaction, along with T4 DNA ligase and ligase buffer, and the reaction incubated at room temperature for two hours.

d. Reaction products were transformed into E.coli, and grown in the presence of ampicillin.

e. Ap^R transformants were isolated, their plasmids prepared using a FastPlasmid® Mini Kit (Eppendorf, Hamburg, Germany), and analyzed by digesting with restriction enzymes AatII and CspCI.

f. Two plasmids were identified, pUC1CspC-1 and pUC1CspC-4, each lacking an AatII site but containing one CspCI recognition site in either of the two possible, opposite orientations. One of these, pUC1CspC-4, was purified on a larger scale, using a Qiagen Plasmid Midi Kit (Qiagen, Valencia, CA) according to the manufacturer's recommendations, for linker insertion at the DraI site.

g. For the DraI site linker, only partial digestion products were desired, therefore digestion, ligation, and DraI site linker components were all added simultaneously.

h. Samples of the reaction were removed and placed on ice after incubation times of 2, 5, 10, 20, 40, and 100 minutes.

i. Reaction samples were transformed into E.coli, plasmids prepared and analyzed as in d. and e. above, digesting with restriction enzymes DraI and CspCI.

j. One plasmid, pUC2CspC, containing two CspCI sites was identified and prepared on a large scale using a Qiagen Plasmid Mega Kit according to the manufacturer's recommendations (Qiagen, Valencia, CA).

Plasmid pUC2CspC was used as the plasmid selection vector for cloning the genes for the CspCI restriction-modification system. Plasmids pUC1CspC-1 and -4 were used as substrates for analysis of the CspCI-cleavage reactions (Example II section b, above).

3. Genomic DNA digestion and library construction

Restriction enzymes ApoI, BamHI, BglII, and Sau3AI were used to individually digest ~10 microgram quantities of *Citrobacter*

sp. 2144 genomic DNA to achieve complete and partial digestions. Following heat-inactivation of the restriction enzymes at 65°C for 15 minutes, the ApoI-digests were ligated to EcoRI-cleaved, CIP-dephosphorylated pUC2CspC vector, and the BamHI-, BglII-, and Sau3AI-digests were ligated to BamHI-cleaved, CIP-dephosphorylated pCspIx2. The ligations, performed overnight with T4 DNA ligase, were then used to transform the endA⁻ *E. coli* host, ER2683 (New England Biolabs, Inc., Beverly, MA), made competent by the $CaCl_2$ method. Several thousand Ampicillin-resistant (Ap^R) transformants were obtained from each ligation. These colonies from each ligation were pooled and amplified in 500ml LB + Ap overnight, and plasmid DNA was prepared from them by CsCl gradient purification to make primary plasmid libraries.

4. Cloning the CspCI genes by methylase-selection

One microgram of each of the primary plasmid libraries was challenged by digestion with ~8 units of CspCI at 37°C for 1 hr. The digestions were transformed back into ER2683 and plated for survivors. Approximately 500 Ap^R survivors arose from the BglII-library, and 5, 29, and 20 from the BamHI-, Sau3AI-, and ApoI-libraries, respectively. Plasmids from BamHI, Sau3AI and ApoI survivors was prepared individually using the Compass Mini Plasmid Kit method, and subjected to CspCI-digestion. 3 of the 20 clones from the ApoI-library were found to be resistant to CspCI, but all those from BamHI- and Sau3AI-libraries were found to be sensitive. The survivors from the BglII-library were pooled and used to prepare a secondary plasmid library. This was challenged

again with CspCI and plated, and among the survivors several additional CspCI-resistant clones were found.

5. Identification of the cspCI-R-M endonuclease-methyltransferase gene, and the cspCI-S specificity gene

The nt sequence of the inserted DNA in the CspCI-resistant plasmid clones was determined by dideoxy automated sequencing. Transposon-insertion into clone ApoI #3, using the GPS-1 System (New England Biolabs, Inc., Beverly, MA), provided the initial substrates for sequencing, and primer-walking was used subsequently, on clones ApoI #3 and #12, and BglII #2 and #17, to finalize the sequence. A total of 4616 bp was determined (Figure 6), within which two complete open reading frames (ORFs) of 1899 bp (nt 1604-3502), and 960 bp (nt 3489-4448) were found (Figure 7). The two ORFs have the same orientation and overlap by 14 bp (Figure 8). Analysis of the ORFs indicated that the larger, termed cspCI-R-M, encodes a combined restriction-and-modification enzyme, R-M-CspCI, and the smaller, termed cspCI-S, encodes a DNA-sequence-specificity protein, S-CspCI (Figure 9). R-M-CspCI is predicated to be 632 aa in length and to have a molecular mass of 70,712 Daltons (or 631 aa and 70,580 Daltons, without the N-terminal fMet). S-CspCI is predicted to be 319 aa in length and to have a molecular mass of 35,267 Daltons (318 aa and 35,136 Daltons without the fMet). Both proteins are necessary for CspCI restriction endonuclease activity.

R-M-CspCI appears to comprise a DNA-cleavage catalytic moiety joined to a DNA-methylation catalytic moiety. Amino acids 2-300, the N-terminal half of R-M-CspCI, more-or-less, are believed to form an endonuclease domain, and to be responsible, primarily, for DNA strand-cleavage activity of CspCI. This section includes the aa sequence motif ...PE-X_{15}-ECK... (aa 57-76), a motif found at the catalytic site of numerous DNA-endonucleases, and likely therefore to be the endonuclease catalytic site of CspCI. Amino acids 301-632 of R-M-CspCI, the C-terminal half of the protein, are believed to form a methyltransferase domain, and to be responsible, primarily, for DNA-modification. This section includes several aa sequence motifs characteristic of the gamma-class of DNA-adenine methyltransferases including ...VLTP... (aa 325-328), ...VLDICAGTGGF... (SEQ ID NO:26) (aa 347-357), and ...NPPY... (aa 435-438). On the basis of this, CspCI is predicted to accomplish modification by methylating adenine residues within its recognition sequence. Symmetry considerations suggest that the bases modified are the second A in the top strand (left sub-sequence), and the only A in the bottom strand (right sub-sequence), thus:

5'...CAA N_5 GTGG...3' -> 5'...CA<u>A</u> N_5 GTGG...3'
(SEQ ID NO:14)
3'...GTT N_5 CACC...5' 3'...GTT N_5 C<u>A</u>CC...5'

R-M-CspCI displays substantial homology to the fused R-M subunit of the BcgI restriction enzyme, and to several similar putative R-M-subunits in Genbank.

S-CspCI also appears to be a fusion protein. In this case, the two sections are similar in sequence and function, and are believed to confer upon CspCI the ability to bind to the two specific components of its recognition sequence. S-CspCI is analogous to, and indeed weakly homologous to, the specificity subunits of type I R-M systems. Amino acids 2-168, the N-terminal half of S-CspCI, more or less, are believed to form one target-recognition domain (TRD), likely the one responsible for binding to the left, 5'-CAA-3', component of the recognition sequence. Amino acids 169-319 are believed to form the other TRD, and likely binds the other, 5'-CCAC-3' component. These two TRDs display considerable homology to each other, and consequently S-CspCI contains several internal repeated sequences. Among these is the proximal repeat INDLF (aa 4-8) and LQDLF (aa 172-176), and the distal repeat PDAYQGVRS (aa 144-152) and PDWDFMEKY (aa 300-308). Similar repeats occur within other specificity proteins, and perhaps mediate in the binding between the S-subunit and R-M-subunit. S-CspCI displays substantial homology to the specificity subunit of BcgI, and to several similar putative specificity subunits in Genbank.

6. Characterization of the cloned CspCI endonuclease

CspCI restriction endonuclease purified according to example 1, above, was subjected to SDS-polyacrylamide gel electrophoresis and found to comprise two proteins of approximately 70 kDa and 35 kDa. High-pressure liquid chromatography of the same sample

demonstrated that the 70kDa and 35kDa proteins occurred in the mass ratio of 1:0.47, implying a molar ratio of 1:1.06. We take this to indicate that CspCI purifies as, and likely is active as, a heterodimer comprising one large subunit (R-M-CspCI) and one small subunit (S-CspCI).

N-terminal sequence analysis of the isolated large subunit indicated that it began with the probable amino acid sequence, ANERKTEELV (SEQ ID NO:27). The initial codons of the CspCI-R-M ORF specify almost the same sequence: MANERKTESLV (SEQ ID NO:28). This result confirms that the large subunit is encoded by the CspCI-R-M ORF; that its translation begins at the predicted ATG at nt 1604; and that the initiating fMet is likely absent in the mature protein. N-terminal analysis of the isolated small subunit indicated that it began with the probable amino acid sequence, PKINDLFHLE (SEQ ID NO:29). The initial codons of the cspCIS ORF specify almost the same sequence: MPKINDLFHLE (SEQ ID NO:30). This result confirms that the small subunit is encoded by the CspCI-S ORF; that its translation begins at the predicted ATG at nt 3489; and that its initiating fMet is also likely absent from the mature protein.

7. Establishing the cleavage site of CspC1

The endonuclease CspCI was found to cleave PhiX174 DNA twice, producing fragments of approximately 3300 bp and 2050 bp. The locations of the cut sites were mapped to approximate positions of nt 1575 and nt 4875 by simultaneously digesting

PhiX174 DNA with CspCI and with additional restriction endonucleases which cleave at known positions, such as PstI, SspI, NciI, and StuI (Figure 1). CspCI did not cut pBR322 DNA or pUC19 DNA. The approximate size of the DNA fragments produced by CspCI digestion of phage lambda DNA (18 kb, 11 kb, 8.3 kb, 5.1 kb, 4.3 kb and 1.8 kb) were entered into the program REBPredictor, which can be accessed at http://tag.neb.com/~vincze/REBpredictor/index.php

REBPredictor uses the algorithm of Gingeras, et al. *Nucl. Acids Res.* 5:4105 (1978), to predict potential recognition sequences by comparing observed fragment sizes with those produced by cleaving the DNA in silico at any given recognition pattern. One predicted potential pattern computed was 5'-CCACNNNNNTTG-3' [SEQ ID NO:31] (or 5'-CAANNNNNGTGG-3' [SEQ ID NO:14] on the complementary strand), which occurs in PhiX174 DNA at positions consistent with the mapping data obtained, i.e. at positions 1563 and 4866. This sequence does not occur in pBR322 or pUC19 DNA. The size of fragments predicted from cleavage at 5'-CAANNNNNGTGG-3' (SEQ ID NO:14) sites in PhiX174, T7 and phage lambda DNAs matched the observed size of fragments from the actual cleavage of these DNAs with CspCI. From these results we conclude that CspCI recognizes the sequence 5'-CAANNNNNGTGG-3' (SEQ ID NO:14).

The positions of cleavage at the CspCI recognition sequence were determined by dideoxy sequencing analysis of

the terminal base sequence obtained from CspCI-cleavage of a suitable DNA substrate, and by comparing the lengths of the CspCI-cleavage products of a labeled DNA to a sequence ladder made from the same primer-template pair (Sanger, et al., *PNAS* 74:5463-5467 (1977); Brown, et al., *J. Mol. Biol.* 140:143-148 (1980)). By the above referenced methods, it was found that CspCI, like several other endonucleases including BcgI, BsaXI, CjeI and HaeIV, cleaves on both sides of its recognition sequence. Our observations suggest that the position of cleavage can vary by one base-pair on either side, being either 5'-N11/N13-CAANNNNNGTGG-N13/N11-3' (SEQ ID NO:32), or 5'-N10/N12-CAANNNNNGTGG-N12/N10-3' (SEQ ID NO:33) or 5'-N10/N12-CAANNNNNGTGG-N13/N11-3' (SEQ ID NO:34) or 5'-N11/N13-CAANNNNNGTGG-N12/N10-3' (SEQ ID NO:35). While not wishing to be limited by theory, we believe the enzyme cuts at a certain distance from the recognition sequence, and that it is the degree of compactness of the DNA within this span that determines whether this results in cutting at 11/13 or 10/12 base pairs.

Example V: Expression of CspCI endonuclease in *E. coli*

The plasmid [pUC19-CspCI-R-M-S ApoI #3] was transferred into ER2683 and plated on Ap^R plates at 37°C overnight. Several individual colonies were inoculated into 50 ml LB+Ap^R and grown at 37°C overnight. All clones expressed CspCI endonuclease activity at $>10^5$ u/g per gram of wet *E. coli* cells. While the

pUC19-CspCI-R-M-S ApoI contains all three domains (cleavage, methylase and specificity moieties) of the endonuclease on a single plasmid for transforming a host cell, it is within the skill of one of ordinary skill in the art to place the cleavage moiety, methylase moiety and specificity moiety on separate plasmids or on a plurality of plasmids in which 2 out of 3 of the domains are present on a single plasmid and the third domain is on a second plasmid.

The strain NEB#1554, ER2683 [pUC19-CspCI-R-M-S ApoI #3] has been deposited under the terms and conditions of the Budapest Treaty with the American Type Culture Collection on March 24, 2004 and received ATCC Accession No. PTA-5887.

Example VI: Engineering Variants of CspCI

CspCI offers a variety of engineering opportunities stemming from its modular organization.

The specificity subunit of CspCI has a duplicated organization that includes a pair of autonomous sequence-selection domains. The domains occur as direct repeats within the linear amino acid sequence, but they adopt reverse orientations in the folded protein to match the anti-parallel organization of double-strand DNA. One domain of S-CspCI is selective for 5'-CAA in dsDNA, and the other for 5'-CCAC; the two domains are separated by about 15 angstroms in the

subunit so that as a whole it recognizes 5'-CAANNNNNGTGG (SEQ ID NO:14) in dsDNA. While not wishing to be limited by theory, it is proposed that actual binding to this sequence involves cooperation between the S-CspCI and the methyltransferase domain of R-M-CspCI, the one sequence-specific, the other non-specific. Alterations introduced into S-CspCI can change the sequence it recognizes in the same ways they have been shown to do in type I R-M systems:

The separation between sequence selection domains and alteration in the length of the non-specific interval in the recognition sequence can be achieved by introducing changes in the 'spacer' region. Examples of such changes include insertions such as small duplications (e.g. to CAA N_6 GTGG [SEQ ID NO:36]) for increased length or deletions to reduce length (e.g. to CAA N_4 GTGG [SEQ ID NO:37]).

Various approaches exemplified below are used to alter the specificity of CspCI.

(a) The recognition sequence of the endonuclease can be altered by tandemly duplicating one of the two specificity domains. In this way, the specificity domain is transformed from recognizing an asymmetric recognition site to recognizing a symmetrical recognition site (e.g. CAA N_5 TTG [SEQ ID NO:38] or CCAC N_5 GTGG [SEQ ID NO:39]). This is

accomplished without physically joining the domains in a single polypeptide chain where dimerization of the tandem repeat can occur spontaneously.

(b) Amino acid changes can be introduced within either domain to alter the sequence selected by that domain, resulting in altered specificity and causing nucleotide discrimination to be diminished (e.g. CAA N_5 GTGR [SEQ ID NO:40]), or lost (e.g. CAA N_5 GTG [SEQ ID NO:41]). Amino acid changes in the S-subunit within the regions flanking the sequence-selection domains are expected to abolish cleavage on both sides of its recognition sequence. The ability of the R-M-subunit to bind to the S-subunit in either orientation can be modified to limit its binding to a single orientation. Accordingly, CspCI, or a variant, may be transformed into an endonuclease that cleaves unilaterally, on only one side of its recognition sequence.

(C) Swaps between the sequence-selection domains of S-CspCI and those of other type IIG enzymes is expected to generate chimeric S-subunits with hybrid specificities. A protein comprising the N-terminus of S-CspCI (recognition sequence CAA N_5 GTGG) (SEQ ID NO:14) and the C-terminus of, for example, S-BcgI (recognition sequence CGA N_5 TGC) (SEQ ID NO:42), when combined with R-M-CspCI may result in an endonuclease that recognizes CAA N_5 TGC (SEQ ID NO:43). For example, N- and C-terminal domains are expected to be

interchangeable to create combinations of two C-terminal domains or two N-terminal domains. In this way, the C-terminal domains of S-CspCI and S-BcgI, together will recognize GCA N_5 GTGG (SEQ ID NO:44). In some Type IIG enzymes, such as HaeIV, AloI, and CjeI, the specificity domain(s) are fused at the C-terminus of the combined R-M-S protein. These can also be swapped into S-CspCI.

Sequence-specificity modules are abundant in nature, occurring both as individual proteins and as domains within composite proteins. Coupling these specificity modules to an endonuclease catalytic site will create endonucleases with new specificities.

Examples of specificity domains from class IIG restriction enzymes that may be used to replace the N- and the C-terminal domains of S-CspCI are as follows:

BcgI (New England Biolabs, Inc., Beverly, MA)
CGANNNNNNTGC (SEQ ID NO:45)

BaeI (New England Biolabs, Inc., Beverly, MA)
ACNNNNGTAYC (SEQ ID NO:46)

BplI (Fermentas GmbH, Vilnius, Lithuania)
GAGNNNNNCTC (SEQ ID NO:47)

CjeI, CCANNNNNNGT (SEQ ID NO:48) from Camylobacter jejuni (Vitor, J.M.B., Morgan, R.D. *Gene* 157: 109-110 (1995)).

AloI (Fermentas GmbH, Vilnius, Lithuania) GAACNNNNNNTCC (SEQ ID NO:49)

HaeIV (Piekarowicz, A., et al. *J. Mol. Biol.* 293: 1055-1065 (1999)) GAYNNNNNRTC (SEQ ID NO:50)

BsaXI (New England Biolabs, Inc., Beverly, MA) ACNNNNNCTCC (SEQ ID NO:51)

In addition to the above, Type I specificity proteins are a rich potential source of specificity-domains for domain-swaps with S-CspCI. The sequence-selection domains of S-CspCI bear some homology to those of the specificity subunits of Type I R-M systems. Hundreds of generally uncharacterized type I S-subunits can be found in Genbank. These proteins interact naturally with Type I modification subunits, which belong to the same gamma-class, of DNA-adenine methyltransferases as R-M-CspCI and can be used as specificity domains for domain swaps.

The C-terminal section of stand-alone gamma-class DNA-adenine methyltransferases is thought to act as a sequence-selection domain, conveying to the otherwise indiscriminate

catalytic site a particular nt sequence to be methylated. These methyltransferases, some solitary, others from Type II and Type IIS R-M systems, abound in nature. Over one hundred have been characterized and many more uncharacterized examples can be found in Genbank. In general, these enzymes recognize continuous nt sequences. Most recognize symmetric sequences 4 to 6 nt in length; others recognize asymmetric sequences of up to 7 nt. These stand-alone methyltransferases also represent a rich potential source of specificity-domains for domain-swaps with S-CspCI. CspCI endonuclease variants with recognition sequences of considerable length could be assembled from these enzymes.

Type I S-proteins interact naturally with Type I modification (M) subunits, forming trimers of composition 2M:1S. These trimers binds specifically to the sequences selected by the S-subunits and subsequently catalyze their methylation. Type I M-subunits are homologous to the C-terminal, methyltransferase, domain of R-M-CspCI, but they lack the N-terminal portion of this protein that forms the endonuclease domain. CspCI can be used to endow endonuclease activity on type I modification enzymes by transferring an endonuclease domain from R-M-CspCI to a type I M-subunit–a 'domain graft'. This will cause the Type I methyltranferase to cleave DNA as well as to modify it.

This experimental approach of grafting the endonuclease domain of R-M-CspCI to the front of a Type I methyltransferase can be applied to other stand-alone methyltransferases to cleave at sequences that originally were only modified. For example, the N-terminus cleavage domain of R-M-CspCI which is a gamma-class DNA adenine methyltransferase can be transferred to other gamma-class DNA adenine methyltransferases.